Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 177 (2007) 969–987
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Infodynamics: Analogical analysis of states of matter and information

Marion G. Ceruti [1], Stuart H. Rubin *

Space and Naval Warfare Systems Center – San Diego, Intelligent Systems, Code 24121, 53360 Hull Street, San Diego, CA 92152-5001, United States

Received 24 January 2006; received in revised form 5 June 2006; accepted 5 July 2006

Abstract

This paper expands and consolidates the use of analogies in thermodynamics to explore concepts in the characterization of information systems. The analogy spans the range of information systems to include databases, knowledge bases and model bases. It includes but is not limited to pressure, expressiveness, temperature, tractability, degrees of order, systems of liquid–liquid equilibrium and disjunction in information-systems integration. By taking advantage of the isomorphism that exists between states of matter and states of information, we can understand new ways to characterize and measure information systems. This paper is the fourth in a series describing new aspects of "infodynamics.".

Keywords: Database; Knowledge base; Infodynamics; Model base; States of matter; Thermodynamics; Expressiveness; Information integration; Metrics; Tractability

1. Introduction

The purpose of this paper is to consolidate and expand the concept of "states of information" as similar to states of matter using analogical reasoning. Differences in states of matter are described with regard to the difficulties in defining each state explicitly. The difficulty in defining the various states of information is seen as a natural consequence of the isomorphism between states of matter and states of information. Taking advantage of this isomorphism, the paper examines the possibility of predicting properties and characteristics of information systems using analogs of well established equations of state and other thermodynamic equations.

Infodynamics is not really a new area of inquiry per se. Other researchers have applied principles of thermodynamics to information systems, particularly in the area of entropy, probability, and reasoning under uncertainty. (See, for example [1,26,32,41,24,40].) Entropy continues to be an active area of research with

* Corresponding author. Tel.: +1 619 553 3554; fax: +1 619 553 1130.
E-mail addresses: marion.ceruti@navy.mil (M.G. Ceruti), stuart.rubin@navy.mil (S.H. Rubin).
[1] Tel.: +1 619 553 4068; fax: +1 619 553 5136.

doi:10.1016/j.ins.2006.07.006

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
15 FEB 2007	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Infodynamics: Analogical analysis of states of matter and information (journal article)	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Marion G. Ceruti, Stuart H. Rubin	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
SSC Pacific San Diego, CA 92152-5001	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Published in Information Sciences, (2007). Volume 177, Issue 4, pp. 969-987. This paper expands and consolidates the use of analogies in thermodynamics to explore concepts in the characterization of information systems. The analogy spans the range of information systems to include databases, knowledge bases and model bases. It includes but is not limited to pressure, expressiveness, temperature, tractability, degrees of order, systems of liquidliquid equilibrium and disjunction in information-systems integration. By taking advantage of the isomorphism that exists between states of matter and states of information, we can understand new ways to characterize and measure information systems. This paper is the fourth in a series describing new aspects of "infodynamics.".

15. SUBJECT TERMS

Database; Knowledge base; Infodynamics; Model base; States of matter; Thermodynamics; Expressiveness; Information integration; Metrics; Tractability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**5**	

an on-line journal since 1999 dedicated to the interdisciplinary approach of entropy in matter and information systems. (See, for example, [23].) Because this aspect of thermodynamics already has received considerable attention in the literature, the present paper does not address entropy, but rather, emphasizes other ways that information systems are similar to systems of matter.

In the first paper in this series on Infodynamics [11], the pressure of a system of molecules was compared to the expressiveness of an information system. Gases were compared to databases and liquids were compared to knowledge bases (KBs) [11]. Temperature also was compared to tractability. In the second paper [12], the dimensions of expressiveness were explored and compared to partial pressures in a gas mixture.

In the third paper [16], the focus shifted to the liquid phase in which the relationship between temperature and tractability was expanded to address the tractability of integrated information systems. Tractability can be conceptualized as the ease of understanding database content, the logic behind its structure and the efficiency of using the database either directly by humans or in applications. Systems of liquid–liquid equilibrium and miscibility were compared to the interaction of data at the interface between two information bases, such as KBs during information-system integration. The relationship between systems of liquid–liquid equilibrium was explored with the idea of application to information systems, their interaction and integration.

Liquids have been compared to knowledge bases (KBs) [11]. Systems of liquid–liquid equilibrium and miscibility are selected for analogical purposes to gain insight into the interaction of data at the interface between two information bases, such as KBs. To date, the relationship between systems of liquid–liquid equilibrium has not been explored extensively for application to information systems, their interaction and integration.

Data integration [9] has been defined clearly in the literature. Data integration occurs when data sets are consistent with each other and free from heterogeneity or conflicts. Data integration represents a tighter coupling between data sets than data aggregation. The three basic levels of data integration are the platform, syntactic and semantic levels [13]. What applies to data integration also applies, even more so in some cases, to knowledge integration. The most challenging level at which to resolve inconsistencies is the semantic level [14].

The paper is organized as follows. Section 2 describes states of matter. Section 3 covers levels of information aggregation. Section 4 describes states of information by analogy to states of matter. Section 5 presents examples of the correspondence between matter and information. Section 6 describes equations of states. Section 7 explores the information analogy of the heat of vaporization. Section 8 covers partial pressures and the information-system analog of expressiveness. Section 9 describes liquid–vapor critical phenomena and their relationship to information systems. Section 10 reviews systems of liquid–liquid equilibrium. Section 11 covers the relationship of liquid mixtures to the integration of information systems. Section 12 explores the concept of a tractability metric that is analogous to temperature. Section 13 explores the concept of information transfer as it relates to diffusion and miscibility. Section 14 describes disjunction metrics and their relationship to ontology and miscibility. Section 15 discusses some key features of an integration as they relate to thermodynamics. Section 16 explores liquid crystals, long-range order and their relationship to information systems. Section 17 discusses the limitations of the methodology. Section 19 suggests future research and applications. Section 19 concludes the paper.

2. States of matter

The three basic states of matter that occur naturally in our environment are gas, liquid and solid. Other states of matter that can occur in a laboratory or in the cosmos include plasma and the dense nuclear material that constitutes neutron stars. This discussion is limited mainly to the naturally occurring states found on earth.

The simplistic definitions for the various states of matter that are offered in introductory science classes and also by Webster are as follows:

- A gas is a substance that has no definite volume or shape; "a fluid (as air) that has neither independent shape nor volume but tends to expand indefinitely" [36].
- A liquid is a substance that has a definite volume but no definite shape; "neither solid nor gaseous; characterized by free movement of the constituent molecules among themselves but without the tendency to separate" [37].

- A solid is a substance that has a definite volume and a definite shape; "neither gaseous nor liquid; a substance that does not flow perceptibly under moderate stress" [38].

Unfortunately, these definitions are insufficient to characterize substances that have properties in between those of liquid and gas, such as dense fluids above the critical temperature. (See, for example, [4–6].) Moreover, they do not characterize accurately substances on the border between liquids and solids, such as liquid crystals. (See, for example, [42,43].) Actually, in the rigorous sense, no clear dividing line exists between liquids and gases, or between liquids and solids. The continuum in the states of matter poses a difficulty in formulating definitions. Ideally, definitions should be crisp so that one can distinguish what an entity is and what it is not. However crisp definitions are not possible in this case because the boundaries between states of matter themselves are fuzzy and not crisp. Fig. 1 illustrates the continuum between gas, liquid and solid, showing variables that either influence or characterize the state of matter.

At the lowest level of granularity, data elements in databases are like individual molecules in gases. The behavior of gases at high temperature and low pressure approaches that of an ideal gas [2]. These gases consist primarily of monomers. In other words, a typical gas at low pressure and high temperature is a collection of single atoms or molecules, each with a trajectory that is separate from that of the other molecules (ignoring collisions with the container wall and with other gaseous species). However, in most physical gases (i.e., not in the theoretical ideal state) a calculable and, in some cases, a measurable fraction of the molecules form clusters of two or more molecules. To form a cluster of N molecules requires an $(N+1)$-way collision. For example, dimers are formed and destroyed by three-way collisions involving three monomers, or a monomer and another dimer. (See, for example, [8].) This clustering effect in a fluid (e.g., gas or liquid) is a precursor to a transition to a more condensed and/or ordered state of matter.

At a higher level of aggregation, knowledge bases are like liquids, which have a great deal of short-range order with respect to the nearest-neighbor internuclear distances. Similarly, knowledge in a knowledge base tends to be clustered in *microtheories*, such as those in the integrated knowledge base. (See, for example, [28,31].)

A microtheory is a set of axioms that pertain to a particular domain and that are consistent within that domain, but are not necessarily correct when used outside of that domain. Microtheories may be detailed enough to be considered to be models, but not all models are microtheories. Some are expressed as systems of equations.

Knowledge bases are analogous to liquids and model bases are analogous to solids. Knowledge-Base Management Systems (KBMSs) are analogous to containers for liquid that have access ports, such as valves and openings. A model base is like a solid – something that can serve as a building material for more complex systems. Domains within the solid are like models in the model base. By analogy, this implies that a large KB with multiple microtheories could be considered to be a form of model base, where the microtheories are the models. It also implies a higher degree of potential usefulness for model bases at a time in the future when we can comprehend and manage them. Fig. 2 shows the relationship between different states of information and expressiveness, tractability and how explicitly the data-relationships are expressed [11].

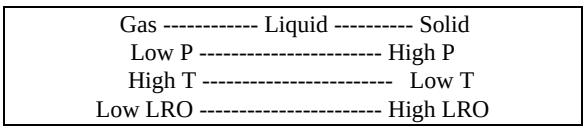

Fig. 1. Effect of variables on states of matter. P = pressure, T = temperature, LRO = long-range order [11].

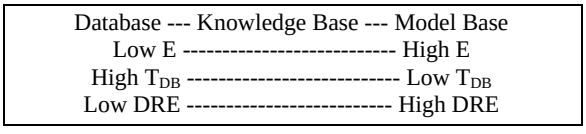

Fig. 2. States of information and their associated variables. E = expressiveness, T = tractability, DRE = data relationship explicitness [11].

The expressiveness–tractability dichotomy [29] in information systems is expected to behave as pressure and temperature in states of matter. A tractable DB is like a gas at high temperatures and low density where intermolecular forces do not provide much influence on the behavior of the gas. Intermolecular forces in matter are analogous to relationships between entities in information systems. Where entities in a DB are disjoint, the DB has few relations. At low P and high T, intermolecular forces do not dominate the behavior of the gas. A DB analogous to this situation is less complex and more tractable [11,12].

In contrast, a KB can be designed and implemented in an information representation that is more expressive than that of a DB. However, a KB with long and complicated rules can be opaque to human comprehension [30]. Thus, as expressiveness increases, tractability decreases [11,12].

3. Levels of information aggregation

The basic unit of information storage in a database is the data element [9]. Similarly, the basic unit of information storage in a knowledge base is the axiom or assertion [10]. An assertion represents information stored at a level of aggregation that is higher than that of a data element. This is because an assertion can involve more than one data elements.

For example, X, A, and B can be stored as data elements in a relational database. (See, for example, [21].) An analysis may be necessary to determine the relationship between these data elements. However, a knowledge base may store the relationship explicitly using a ternary predicate. For example, the assertion could be that X is between A and B. (For probabilistic knowledge bases, such as Bayesian networks, the information aggregation issue is more complicated as the knowledge is stored in the network structure and in the conditional probability table. See, for example, [15].)

A model base is a repository of models. Models represent a state of information aggregation that is at a higher level than that of knowledge. Models show the relationship between knowledge in an explicit manner, just as knowledge expresses the relationship between data elements explicitly. This relationship often is expressed as an equation or a group of equations, a computer program that captures an algorithm or heuristics, or in a variety of other ways depending on how the models are to be used.

What comes after model base in the DB–KB–MB progression? What happens when you aggregate models? The periodic table of the elements enables chemists to predict the properties of elements that are not yet discovered. Similarly, one can predict using analogical reasoning the next member in the DB–KB–MB series. This should be an aggregation of models constructed in a useful manner to produce, what for lack of a better term may be called a *wisdom base* (WB).

Information aggregation, when accomplished correctly to build an information system, is like an aggregation of atoms and molecules used to form a specific and definite physical structure. Just as a useful, solid object with specific properties (such as a tool) will not consist of just any random or arbitrary aggregation of molecules, we need an exact, specific structure in an information system for that system to be useful for its intended purpose. Similarly, any arbitrary aggregate of data will not necessarily constitute a knowledge base and any arbitrary aggregate of knowledge, especially where disjoint, will not be likely to constitute a model base.

4. States of information

Databases (DBs), knowledge bases (KBs) and model bases (MBs) are information repositories in which information is stored in progressively higher levels of aggregation and complexity [10]. A *database* is a state of information that consists of facts or figures structured according to a model that allows knowledge to be stored implicitly and from which conclusions can be inferred [10,11]. At the lowest level of granularity, data elements in databases are like individual molecules in gases. The behavior of gases at high temperature and low pressure approaches that of an ideal gas in which molecules behave independently [2].

A knowledge base is of two types. A type-one *knowledge base* is a state of information that consists of a collection of rules, axioms or assertions structured according to an ontology and a knowledge representation that allows knowledge to be stored explicitly, and from which conclusions can be drawn using an inference engine [12]. A type-two knowledge base is a structured acyclic graph, such as a Bayesian network that stores

knowledge in its structure and in its associated conditional probability table. Most of the discussion on knowledge bases in this paper is limited to type-one knowledge bases. A *model base* is a state of information that consists of models, in which knowledge is aggregated either implicitly or explicitly [12]. A model base is structured according to a system that allows interactions and relationships between models to be exploited and from which conclusions can be inferred using software tools. If the model itself is treated as a representational formalism, then the distinction between types I and II knowledge bases blurs. This is similar to the principle of duality in physics.

In databases the information is referential, in knowledge bases it is inferential, and in model bases, the information is experiential. Although calculations and recursion can be accomplished through database queries, the primary function of database is to serve as a reference. Similarly although a knowledge base can be used as a reference by programming look-up tables into axiom format the strength and power of a knowledge base when combined with an inference engine lies in its capability for inference. Finally, models can be understood by applying them to tasks versus through theoretical explanation. This is especially true of probabilistic networks. Thus, models provide experience just as databases provide reference and knowledge bases provide inference.

5. Examples

X, A, and B can be stored as data elements in a relational database. (See, for example, [21].) An analysis may be necessary to determine the relationship between these data elements. However, a knowledge base may store the relationship explicitly using a ternary predicate. For example, the assertion could be that X is between A and B [11]. For probabilistic knowledge bases, such as Bayesian networks, the information aggregation issue is more complicated as the knowledge is stored in the network structure and in the conditional probability table [15].

To a first approximation, the states of information described above are isomorphic to states of matter. Table 1 summarizes the comparison between the domains of matter and information. The information contained in DBs, KBs, and MBs is in different states, or ''states of information''. The same information can occupy different states in different information bases, just as molecules occupy different states of matter, depending on temperature and pressure.

The state that the information occupies depends at least on the type of information base that stores the data, the level of tractability of the information, and the level of expressiveness that the information management system enables. For example, to express in a relational database the relationship between the lengths of ships and their beams, the database administrator would create a table with at least the following attributes (probably more), ship name, hull number, length and beam. The next step would be to fill the table with data on actual ships. Upon inspection, it would be obvious that a ship's length always exceeds its beam. This fact is stored implicitly in the relational database and can be made more explicit by issuing the appropriate query [11]. A database is a kind of knowledge base that allows a specific type of inference [10,29].

In contrast, to express the length–width relationship in a knowledge base, a knowledge engineer would write an explicit assertion stating in the language of the knowledge-base representation the following axiom: ''Always true: Length.ship > beam.ship''. In a model base, this fact might be incorporated into a model that a naval architect could use to design a ship with a hull that produces less drag than ships available today. The length–beam relationship would be part of a model that describes the basic hull configuration. An equation would relate the two as independent variables that determine, among other variables, the drag, degree of laminar flow, and maximum hull speed. From a model base, one could understand in terms of water resistance, why a ship is always longer than it is wide [11].

Data stored in databases the relations of which are in at least first normal form are analogous to molecules in the gas phase. Even the terminology of information systems here is similar to that of chemistry (e.g., element, atomic, etc.). The term *data element* implies that the information at that level cannot be broken down further and thus possesses the property of *atomicity*. Databases and their management systems are analogous to gas-handling systems with manifolds, gauges, valves, and gas cylinders. (See, for example, [7].) These aggregates of molecules in the gas phase are analogous to correlated aggregates of data from database queries, such as data in relations. Just as a dimer consists of two molecules that have the roughly same translational trajectory between collisions, data aggregates in databases can be formed by ad hoc join queries that bring

Table 1
Comparison of variables and observed phenomena for the domains of matter and information [11]

Variable or observation	Matter	Information
Smallest unit	Atom or element	Data element
State variable	Pressure; chemical potential	Expressiveness [11]
State variable	Temperature	Tractability [11]
Basic mass unit	Atomic or molecular weight	Importance or priority of data element for maintenance, updates & integration purposes. Assigned by database administrator per [17]
Phenomenon that correlates the behavior of entities	Intermolecular forces	Relationships between entities; semantic distance between concepts in an ontology; interdependence of variables
State of lowest order, not condensed	Gas	Database [11]
State of intermediate order, condensed fluid	Liquid	Knowledge base [11]
State of high order along multiple dimensions; state of high potential usefulness as building material for tools	Solid	Model base
State of extreme aggregation, density and complexity	Neutron stars	Wisdom base
Process that initiates gas–liquid phase transition; precursor to state of higher aggregation, complexity, and local order	Nucleation in gases	Table creation, formation of semantically heterogeneous groups [14]
Process that initiates liquid–solid phase transition; precursor to state of higher aggregation, complexity, and long-range order	Crystallization, or seeding in liquids	Cluster generation in ontologies and in knowledge bases; Seed concept identification [30,31]
Intermediate state between gas and liquid	Critical mixture, dense fluid	Storing data in a knowledge base or storing knowledge in a database
Intermediate state between liquid and solid	Liquid crystals	Large, expressive knowledge bases that contain many microtheories or clusters [28]
Integration mechanism	Emulsifier	Ontology [16]
Tendency to resist merging	Immiscibility	Disjunction [16]
Translational motion	Diffusion	Information transfer [16]

together data from two or more tables to satisfy what is frequently a specific, immediate, and temporary requirement.

Proceeding to a higher level of aggregation, knowledge bases are like liquids, which have a great deal of short-range order with respect to the nearest-neighbor internuclear distances. Similarly, knowledge in a knowledge base tends to be clustered in *microtheories*, such as those in integrated knowledge bases. (See, for example, [28,31].) A microtheory is a set of axioms that pertain to a particular domain and are consistent within that domain, but are not necessarily correct when used outside of that domain.

Interestingly, in a crystalline solid, a ''domain'' is a region of the material in which long-range order persists, and in which the location of one atom or molecule can be predicted with a high degree of accuracy given the locations of other molecules. This is not the case for prediction concerning adjacent domains, where the long-range order proceeds along an access with a different orientation. One cannot predict the position of an atom across multiple domains with the same degree of certainty as is possible within a single domain.

6. Equations of state

Just as states of matter are not well defined, databases, knowledge bases, and model bases are not well-defined concepts in general [10]. This becomes readily apparent when comparing and contrasting the states of information. As long as the molecules under consideration are located far from phase interfaces, the states of matter look better defined under some circumstances. Most of the difficulty with finding crisp definitions for states of both matter and information arises when attempting to compare and contrast the different states at their boundaries. The domain isomorphism between states of matter and states of information, which is summarized in Table 1, enables us to understand why we have such difficulty in formulating crisp definitions for terms like *database*, *knowledge base*, and *model base* in simple, succinct terms. (See, for example, [10].) Both

state sets consist of members with fuzzy boundaries. Furthermore, cross-domain analogies are usually not defining, but rather serve as heuristics guiding the evolution of one ontology from another.

So far, no one has developed an equation of state similar to the ideal gas law for a database. The following considerations will be useful to take the first step in that direction. The ideal gas law is given by equation [2]:

$$PV = NkT, \tag{1}$$

where P is pressure in atmospheres, V is volume in liters, N is the number of molecules (or atoms in the case of noble gases) and T is the absolute temperature. The constant of proportionality, k, is Boltzmann's constant, which is equal numerically to 1.3623×10^{-21} l atm./molecule/deg. Ideal gases are assumed to consist of molecules that occupy no space and have no intermolecular forces.

Using this formula, consider an equation of state for a database. Suppose we redefine N as the number of atomic data elements. We assume that T_{DB} is a measure of tractability (analogous to temperature) and E is a measure of expressiveness. E is analogous to P in a gas system (i.e., $P_{\mathrm{DB}} = E$). So T_{DB} and E in a database system are analogous to T and P, respectively, in a gas system. The choice of variables is appropriate for two reasons.

First, two definitions of the verb, *express*, are "to force out by pressure" and "to subject to pressure so as to extract something" [39]. Whereas this is not the same definition of "express" that ordinarily would be associated with an information system, both information expressiveness and expression through pressure [39] are about bringing something outside (in a form in which it can be observed, understood and used) that previously was inside (in a form less observable and useful).

Thus, expressiveness, E, in a database system is an appropriate analog for pressure, P, in a gas system. It is reasonable to assume that the expressiveness of an information system would be directly proportional to the amount of distinct and non-redundant information in it, although N is by no means the only factor to determine expressiveness [11]. E represents the richness of detailed ideas and concepts implicit in the data and the ease with which they can be extracted. Issuing a query in a database is like opening a valve in a manifold that holds fluid under pressure, ignoring the decrease in pressure that results from the change the amount of material. (See Section 17.)

Second, P and T affect the volume of a gas in opposite directions. At constant N, an increase in P will decrease V whereas an increase in T will increase V. Similarly, E and T_{DB} work in opposite directions in a database with the same number of data elements. As E increases at constant N, T_{DB} decreases. E and T_{DB} were selected to account for the well-documented tradeoff between expressiveness and tractability that is like a reciprocal relationship [29].

V_{DB} is a volume-like entity that changes as E and T_{DB} change at constant N. V_{DB} is related to the scope, S, of the database, i.e., the number of topics and level of detail of each topic:

$$V_{\mathrm{DB}} = S. \tag{2}$$

Thus an equation of state for a database analogous the ideal gas law would look something like:

$$ES = Nk_{\mathrm{DB}}T_{\mathrm{DB}}. \tag{3}$$

As the scope of the database increases at constant N and T_{DB}, the expressiveness, E decreases because in this case, the information in the database must be spread out over a larger scope with less expressive detail in any one specific area. If the scope, S, and number of data elements, N remain constant, as the expressiveness E increases the tractability, T_{DB} also increases. This is intuitive because to increase the expressiveness, one may need to change in the database structure through, for example, normalization. This could lead to less confusion about the entities that data elements describe. Alternately, in an effort to increase expressiveness without increasing the size or scope of the database, the data themselves may have to be expressed more concisely and clearly, thus increasing tractability, T_{DB}.

Solving for k_{DB} in Eq. (3), one arrives at an expression for k_{DB}, which is like Boltzmann's constant for database systems:

$$k_{\mathrm{DB}} = ES/NT_{\mathrm{DB}}. \tag{4}$$

Whereas the ideal gas law is useful for understanding certain basic behavior of gases, in fact, no physically observable gas is an ideal gas. Similarly, whereas an ideal equation of state for information systems analogous to the ideal gas law may be of some theoretical value, it is not of much practical use for some systems because most, if not all, large databases and knowledge bases are replete with relationships between the entities. These relationships are like intermolecular forces in gases that couple the behavior of the various entities, linking many interdependent variables together. Such linkage is very similar in some ways to the coupling between molecules that occurs in viscous fluids. Here, momentum transfers easily from one species to the next, thereby frustrating any hope of being able to treat most modern information systems with the simplicity of an ideal-gas-like equation of state. Still, Eq. (4) invites us to examine the issue of metrics. (Sections 8 and 12.)

The next simplest equation of state after the ideal gas law is the van der Waals equation (5) where Eq. (6) defines the molar volume, R is the gas constant, and A is Avogadro's number, which is 6.023×10^{23} molecules/g molecular weight or mole [3]. In chemical systems, Avogadro's, A, number is equal to the number of atoms in a gram of hydrogen. It is a scaling factor between microscopic and macroscopic quantities of matter [11]. Constants, "a" and "b," represent corrections for molecular size and intermolecular forces respectively, which differ for each gas.

$$P = RT/(\underline{V} - b) - a/\underline{V}^2, \tag{5}$$

$$\underline{V} = (AV)/N, \tag{6}$$

$$R = Ak. \tag{7}$$

Eqs. (8) and (9) give the database-systems analog of (5).

$$E = (A_{\mathrm{DB}} k_{\mathrm{DB}} T_{\mathrm{DB}})/(\underline{V}_{\mathrm{DB}} - b_{\mathrm{DB}}) - a_{\mathrm{DB}}/(\underline{V}_{\mathrm{DB}})^2, \tag{8}$$

$$\underline{V}_{\mathrm{DB}} = SA_{\mathrm{DB}}/N. \tag{9}$$

A_{DB} is like Avogadro's number in that it could be related to scalability in databases. A_{DB} will not, however, have exactly the same meaning in the information context that Avogadro's number has in the material context.

Van der Waals constant, a, corrects for molecular size [11]. The constant, a_{DB}, is the information-system analog of the van der Waals constant that represents the increase in expressiveness of a database with comment or text fields that allow for declarative information to be included in database format. Here, the size of the field is analogous to atomic or molecular size.

Similarly, b_{DB} is the information analog to the van der Waals constant that corrects for intermolecular forces [11], which usually are attractive forces at long range. b_{DB} is related to the degree to which relationships between data elements have been made explicit. Whereas no metric for b_{DB} has been developed, a low b_{DB} would indicate the presence of implicit or latent correlating relationships between data elements that have not be made explicit. In a database characterized mainly by disjoint data elements b_{DB} would be near zero, like the ideal-gas case in which no forces are assumed to act between molecules. For example, dependence is a form of correlation. If data elements were shown to depend on each another, that would tend to increase b_{DB}.

As b_{DB} increases, E also increases, subject to the constraint that b_{DB} must remain small compared to $\underline{V}_{\mathrm{DB}}$ (and they can never be equal). Providing better documentation in the database about the relationships between data elements can be conceptualized as an increase in b_{DB}. This also leads to better expressiveness of the database, as the database complexity approaches that of a knowledge base, where relationships are more explicit. The process of deriving new data using relationships between existing data is very similar to the generation of features in a database to aid in the knowledge-discovery process. (See, for example, [34].)

7. Heat of vaporization

As data relationships are characterized, the database approaches a knowledge base in which all information can be expressed as declarative statements or axioms. This suggests the possibility of a phase transition. For example, one can define the quantity, Q_{IV}, as the "work of database conversion", which is the direct analog of

the thermodynamic quantity, Q_{vap}, or heat of vaporization. For a van der Waals gas, Eq. (10) defines Q_{vap} as follows [4]:

$$Q_{vap} = a/b. \tag{10}$$

To a first approximation, for information systems, the work necessary to convert information between knowledge base and database representations (e.g., KBMS ↔ DBMS) is directly proportional to the number of latent relationships in the data that need to be made explicit. Q_{IV} also is inversely proportional to the degree of "disjointedness" of the information. Eq. (11) summarize the relationship and can be viewed as a measure of the complexity of the information–representation conversion.

$$Q_{IV} = a_{DB}/b_{DB}. \tag{11}$$

For high Q_{IV}, many relationships exist between data elements that necessitate explicit declarations in a corresponding knowledge base. For low Q_{IV}, the task of conversion is simpler either because relationships have been made explicit or because fewer relationships exist, in which case the domains of related variables or microtheories can be handled separately from each other. The concepts and usage of both a_{DB} and b_{DB} need to be refined. Moreover, a way to measure overall disjunction in an information system is required.

8. Partial pressures and expressiveness

Expressiveness can occur along multiple dimensions, which, to a first approximation, can be conceptualized as additive like partial pressures. An information system can be expressive in the following ways [12]:

- e_1 – To a first approximation N, the number of data elements in an information system, could serve as a reasonable estimate of e_1.
- e_2 – An information system is expressive if it supports high-resolution concepts by allowing the user to distinguish between entities when the differences are very small, i.e., the ontology is very rich because it allows for many fine gradations of the same or similar concepts. For example, a paint manufacturer may have many different names for different shades of blue. Here, the dimension of expressiveness, e_2, could be estimated by a quantification of the fan-out of entities at various levels in the ontology. It also could be characterized by comparing several different information bases and rank ordering them according to the magnitude of the just-noticeable differences that can be expressed.
- e_3 – An information system can provide multiple synonyms for the same entity, thus increasing the probability that the system can support users from different backgrounds where different terminology is used to express the same concept. Here, the dimension of expressiveness is synonomy. A simple way to measure e_3 is to count synonyms.
- e_4 – It can handle multiple query types, such queries that include negation, counterfactuals, and uncertainty. An estimate of e_4 is to count the number of query types that the information system supports.

Dalton's law of partial pressures is stated as follows:

$$P = p_1 + p_2 + \cdots + p_n, \tag{12}$$

where $p_1, \ldots, p_n$ represent the partial pressure of each gas in the system and P is the total pressure.

Similarly, the total expressiveness of an information system can be considered to be the sum of the expressiveness along each dimension of expressiveness:

$$E = c_1 e_1 + c_2 e_2 + \cdots + c_n e_n, \tag{13}$$

where $e_1, \ldots, e_n$ represent the partial measures of expressiveness along each dimension that is present in the system, three of which are described above. Constants, $c_1, \ldots, c_n$ are included in (13) to make the equation more flexible in that some dimensions of expressiveness may be more important than others, depending on the application. In the absence of any other information, each of these constants can be set equal to 1. Eq. (13) holds as long as the dimensions of expressiveness are orthogonal and all e_i are obtained by counting entities.

9. Critical phenomena and nucleation

The liquid–vapor critical point is the temperature and pressure at which the interface between a liquid and the vapor of that substance over the liquid disappears [2,4]. To observe critical phenomena experimentally, partially fill an evacuated pressure vessel with a liquid at room temperature and seal the vessel. Heat the vessel until the interface between the liquid and the gas above the liquid vanishes. The state of matter is not defined clearly for a substance with a temperature above its critical point. It is a dense fluid that clearly is not a solid. However, whether it is a liquid or a gas cannot be determined.

The information-system analog of critical phenomena is the database–knowledge base transition. A database is a kind of knowledge base that allows for a specific type of inference [10]. Databases can be constructed to store axioms and assertions as text fields. They also can contain the conditional probabilities associated with Bayesian networks. Knowledge bases can be constructed to contain assertions that might also be expressed very efficiently in tabular format. Under some circumstances, it may not be any easier to distinguish a database from a knowledge base than it is to separate liquid from vapor above the critical temperature, unless one examines the information representation and the query methods.

Information grouping can be compared to nucleation in matter. This area needs to be explored further. Data grouping in databases [14], clustering in data streams [27] and axiom clustering [31] in knowledge bases are analogous to nucleation in gases and crystallization in liquids respectively because they initiate phase transitions to states of information with longer-range order and correlation among information entities. This is because these grouping techniques bring together data or knowledge in which the relationships between data elements or axioms link the elements together in the cluster or group in a manner analogous to the way in which intermolecular forces hold atoms or molecules together in condensed phases of matter.

10. Systems of liquid–liquid equilibrium

Some pairs of liquids are immiscible with each other under certain conditions that depend on temperature and composition. They can become partly miscible or totally miscible if the temperature or composition changes.

The liquid–liquid critical point is the temperature, T_c, and composition (i.e., mole fraction, x_c,) at which the liquid–liquid interface at equilibrium disappears and the two liquids become miscible with each other [2]. A phase diagram specific to each liquid–liquid pair describes the behavior of the liquid with respect to critical temperature and composition. In many systems of liquids, T_c occurs at the maximum, and in some systems, T_c will occur at the minimum of the curve.

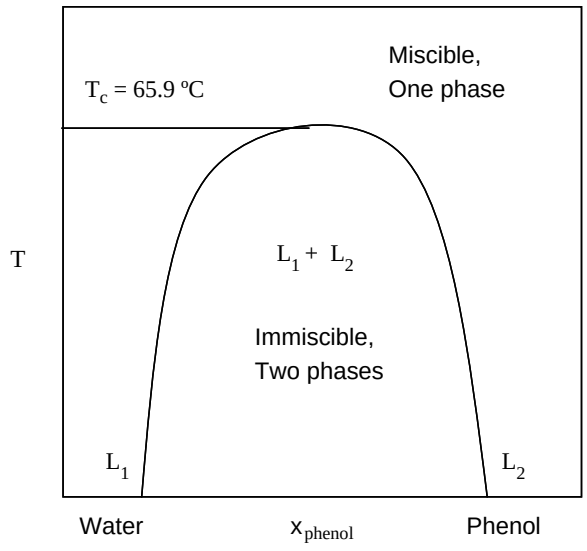

Fig. 3. Liquid–liquid phase diagram for water and phenol [2]. ("x_{phenol}" is the mole fraction of phenol).

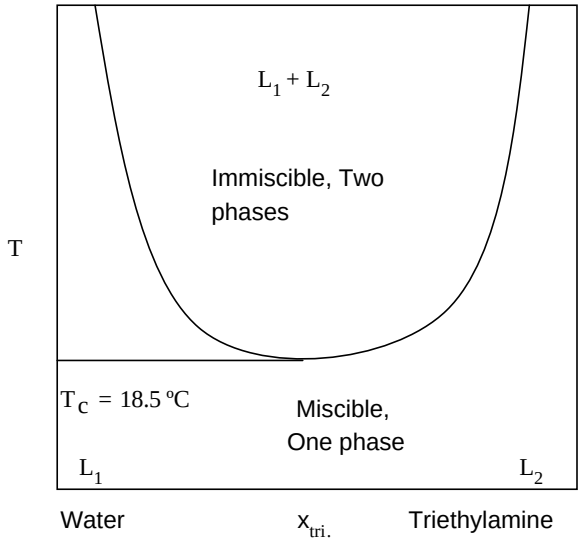

Fig. 4. Liquid–liquid phase diagram for water and triethylamine [3]. ("x_{tri}" = mole fraction triethylamine).

Figs. 3 and 4 illustrate the two cases, respectively. The critical temperature can be either an upper consolute temperature as in Fig. 3 or a lower consolute temperature as in Fig. 4 [2]. On one side of the curve ($L_1 + L_2$), the liquids are immiscible and an interface forms between the two, whereas on the other side of the curve, the liquids are miscible and exist as a single phase.

The thermodynamics of the liquid pair is given by Eq. (14).

$$u_a - u_a^0 = RT \ln x_a, \tag{14}$$

where x_a is mole fraction of liquid "a" in the system, u_a^0 is the chemical potential of the pure liquid, and u_a is that of the liquid in equilibrium with the other liquid. In liquids, the variable, u, can be conceptualized as a gas pressure-like quantity, similar to vapor pressure. R is the constant of proportionality that was determined experimentally.

11. Relation of liquid–liquid mixtures to information systems

Table 1 includes a comparison between the domains of matter and information when comparing integration between two information bases (DB or KB) to a system of two liquids.

Consider two KBs, "1" and "2" that are proposed for integration. Eq. (15) is the information-system analog of Eq. (14). In (15) E_1^0 is the expressiveness or "information potential" of KB_1 in the integrated state, E_1 is the expressiveness or "information potential" of KB_1 in the stand-alone state, and T_{kb} is the tractability of the information system. x_{1kb} is the fraction of information contributed from KB_1. R_i, the constant of proportionality like R in gases, will need to be determined experimentally:

$$E_1 - E_1^0 = R_i T_{kb} \ln x_{1kb}. \tag{15}$$

Measures of E have been described [12] and x can be approximated by counting attributes in databases (DBs) or axioms in KBs. Metrics for T_{kb} are explored in the next section.

Partial miscibility of two liquids is like two KBs that have been integrated at some levels but not at all levels. In principal, this applies to mixtures of multiple components and the results can be generalized to systems of multiple KBs. A method to measure disjunction in KBs needs to be developed in analogy with the immiscibility of liquids. Such a metric will need to be generalized to include heterogeneous information systems types (e.g., systems that include both DBs and KBs) as well as information systems that include multiple components of the same type.

T_c is not easy to predict or calculate from other characteristics of liquids, such as boiling points, freezing points and molecular structure. Similarly, it is not envisioned that T_{ci}, the critical tractability of information

integration, will be easy to predict or calculate theoretically. However, using approximate metrics for tractability a method to measure T_{ci} experimentally could be developed in analogy with T_c for liquid systems.

12. Toward tractability metrics

Molecular motion gives rise to temperature T. In systems of molecules, different kinds of motion give rise to various contributions to T. More specifically, heat is partitioned among orthogonal motion types that give rise to the rotational, vibrational, and translation temperatures. The motion types are derived from the orthogonal degrees of freedom of the atoms in each molecule. Each type of motion makes a separate contribution, each of which can be calculated theoretically. However, experimentally using direct measurements, we observe an overall T resulting from the contributions of all degrees of freedom. (Indirect spectroscopic means must be used to determine the contributions to T from vibration and rotation, given sufficient spectral resolution.) Similarly, a metric is needed for the information tractability in information systems analogous to temperature.

Multiple aspects of tractability are possible. The overall tractability can include contributions from all aspects. For example, it can include contributions from the various independent levels of integration (e.g., platform level, syntactic level and semantic level) [14]. These levels, which progress from the coarse grain (platform) to the fine grain (semantic), may be compared in some ways respectively to the orthogonal motion types, such as vibration, rotation and translation (also in decreasing order of granularity). Information system tractability with respect to integration means that obstacles to integration at the various levels of integration are overcome efficiently. Metrics are needed for each of these three main integration levels to determine in more detail their contribution to the overall tractability.

Other aspects of tractability also may contribute to overall tractability, such as the aspect from the point of view of the engineer who must integrate the information (efficiency of integration) and another from the point of view of the user (ease of use). From the user's perspective, tractability of an information system relates to the steepness of the learning curve in understanding the information in the system and in using the system to meet mission requirements. Consider Eq. (16) as a formula for information-system tractability in which T_1 is the contribution to the tractability from integration at the platform level; T_2, the tractability at the syntactic level; and T_3, the tractability at the semantic level. T_1, T_2, and T_3 pertain to the level of effort on the part of the engineer. T_4 can be added to represent the tractability contribution from the user's viewpoint. The c_n constants are weighting factors that each can be set arbitrarily to 1 unless there is some a priori reason for making them unequal.

$$T_{kb} = c_1 T_1 + c_2 T_2 + \cdots c_n T_n. \tag{16}$$

As in the case of molecular systems in which the individual contribution of the various components to the overall temperature are difficult to measure separately and directly, the contributions of the various aspects of tractability also are not measured easily. However, they can be estimated. For example, T_1, T_2, etc. can be estimated separately using a scale of, say, 1–10. For example, T_1 is the tractability of the platform-level of integration, which includes basic hardware, network connectivity and protocol, operating systems, and transaction management [14]. To get a full score of 10, no aspect of platform connectivity would be allowed to decrease the efficiency or throughput of the system.

T_2 is the tractability at the syntactic level of integration, which includes data structures, languages (e.g. SQL, KQML) and constraints [13]. T_3 is the tractability at the semantic level, which includes data-element naming conventions, definitions, units, levels of granularity, precision [13], and ontology placement. T_2 and T_3 also can be estimated in a similar manner to that of T_1 including the various contribution to each T_n from data structures, languages, semantic inconsistencies, etc.

T_4 is the tractability from the point of view of the user. This includes ease of use, understanding, and task-reduction time. The reliability of each platform also is a consideration as it relates to the perceived tractability of integrated information. Using this system to estimate T_{kb}, the T_{kb} of one system can be compared to that of another provided the individual components of T_{kb} are estimated using the same criteria. Absolute values of T_{kb} may not be as useful or as meaningful.

13. Diffusion and information transfer

One way to conceptualize the tractability is to note that two KBs at low tractability are like a two-phase liquid–liquid system at equilibrium with an interface that allows little transfer to occur. Diffusion in liquid–liquid systems can be compared to information transfer in information systems of multiple components. As indicated in Table 1, the disjunction of information systems is analogous to immiscibility in liquid–liquid equilibrium systems. Diffusion across the boundary between liquids is analogous to information transfer between one information-system component and another.

For information systems there will be a level of tractability (given a certain relative amount of information in each system) at which integration becomes very efficient. This gives rise to a critical tractability, Tc_{kb}, analogous to the critical temperature in liquid systems, T_c. At Tc_{kb}, diffusion-like information interoperability can occur readily between the two components and the interface between them can be made transparent to the user, just as the interface between two liquids vanishes at the critical temperature and composition as previously discussed.

If the interface allows little meaningful information transfer, few axioms from KB_1 can be used in KB_2. Tractability is low here. The analog is the H_2O–phenol liquid–liquid system depicted in Fig. 3 that results in two phases at some concentrations below the critical temperature. Most of the time, information systems are expected to behave more like the H_2O–phenol phase diagram in Fig. 3 than the water–triethylamine phase diagram depicted in Fig. 4. This is because in general, as tractability increases, the probability of a two-phase system decreases and the information "miscibility" or the efficiency of information transfer increases.

A method needs to be developed to determine if two KBs are miscible or in two phases. This is like asking if the KBs are disjoint, and at what level in the underlying ontologies do they have concepts in common.

The example is given of two KBs as analogous to a system of two liquids, either miscible or immiscible, depending on the degree of their molecular polarity (like the degree of disjunction in information systems). In the case of a two-phase system of liquids, molecules of both types are exchanged across the liquid–liquid interface so that some of liquid A dissolves in the B phase and some of liquid B dissolves in the A phase.

Even partial miscibility can result in a two-phase system when A becomes saturated in B or vice versa. Beyond the saturation mole fraction at constant temperature, increments of either component will not mix but will result in a second phase appearing with an interface between the two phases. The saturation mole fraction depends on temperature and measuring it at various temperatures gives rise to curves such as those depicted in Figs. 1 and 2.

14. Disjunction metrics, ontology and miscibility

To a first approximation, disjunction in an information system is analogous to immiscibility in a multi-liquid system. Other factors can produce a "two-phase" KB_1–KB_2 system if the knowledge representations are very different. No axiom in A will appear to form useful clusters with the axioms in B. Today, this occurs as microtheories in large KBs in which the domains are disjoint. However, if tractability is increased by converting information from A into the knowledge representation of B, the two-phase system may become a one-phase system consisting of A and B as miscible KBs like the miscible liquids.

Methods have been suggested to characterize, estimate, and eventually measure disjunction in information systems [18], which is the analog of immiscibility of liquids. For example, consider two KBs, KB_1 and KB_2. The higher (more general) level the ontology or KB structure that is necessary to find axioms or concepts in common with another KB, the more disjoint (i.e., orthogonal or mutually random) two KBs are from each other. One can count the levels starting from the leaves (most specific instance level) calling this level zero. The next level is 1, etc. Therefore, one could say, for example, that an axiom from KB_1 and another one from KB_2 are disjoint at the $(3, 5)$ level where 3 represents the level of generality/specificity in the ontology in KB_1 that corresponds to level 5 in KB_2. The higher the numbers, the more disjoint the axiom in KB_1 is from the axiom in KB_2. This disjunction concept is captured in Eqs. (17) and (18), which apply to the single group of three axioms from the example described above:

$$Dj(\mathrm{KB}_1(a_3), \mathrm{KB}_2(a_5)) = (3,5), \tag{17}$$

$$Dj(\mathrm{KB}_1(a_3), \mathrm{KB}_3(a_8)) = (3,8). \tag{18}$$

Eqs. (17) and (18) are examples of the disjunction metric, $Dj(x,y)$, that can be used to compare axioms in KBs. Eqs. (17) and (18) can be used to compare the degree of disjunction between pairs of axioms from different databases. To use this metric, the ontology that pertains to each KB must be sufficiently complete to locate the corresponding levels in the ontologies of the different KBs. Disjunction also is related to randomization in information systems. (See, for example, [19].)

Another way to express the disjunction metric is with Eqs. (19) and (20). Eq. (19) states that a concept at level 3 of ontology for KB_1 is equivalent to a corresponding concept at level 5 of ontology for KB_2. Eq. (20) is the analog of (19) in the case of KB_1 and KB_3.

$$(\mathrm{KB}_1(a_3)) = (\mathrm{KB}_2(a_5)), \tag{19}$$

$$(\mathrm{KB}_1(a_3)) = (\mathrm{KB}_3(a_8)). \tag{20}$$

Given (17) and (18), we can also write (21).

$$Dj(\mathrm{KB}_2(a_5), \mathrm{KB}_3(a_8)) = (5,8). \tag{21}$$

Similarly, given (19) and (20), we can also write (22):

$$(\mathrm{KB}_2(a_5)) = (\mathrm{KB}_3(a_8)). \tag{22}$$

Moreover, one can sum the axioms or concepts from one KB at level x that occur at level y in another KB and divide by the total number of axioms at that level in each KB to calculate an overall disjunction metric, $Dj(1,x,2,y)$ at the (x,y) level of comparison. Eqs. (23) and (24) express disjunction about an aggregate of axioms or concepts. Integers, k and m are the total number of axioms or concepts at levels x in KB_1 and y in KB_2:

$$\sum Dj(\mathrm{KB}_1(a_x)), (\mathrm{KB}_2(a_y)) = \sum (x,y), \tag{23}$$

$$Dj(1,x,2,y) = \sum (x/k, y/m). \tag{24}$$

The usefulness of these disjunction metrics will increase when a more standardized way to organize an ontology is developed.

An example of partial miscibility in liquids is to dissolve small amount, say 5% of phenol in water and still maintain a one-phase system, as shown in Fig. 3. In analogy with partial miscibility, if only a small amount of information from one source (e.g., DB or KB) is integrated with another larger information base, this can be approached in a tractable way just by performing exhaustive searches and comparisons.

When the sources are of comparable size and both are large, it becomes more difficult, if not impossible to integrate these sources at all levels by manual and exhaustive means as this method of integration is not scalable. This situation corresponds to the two-phase side ($L_1 + L_2$) of the critical temperature in a liquid–liquid equilibrium system. Within this boundary, which corresponds to the area below the curve in Fig. 3 and the area above the curve in Fig. 4, liquids do not mix well with each other and two liquid phases result. In the information analog, an information system will be difficult to integrate in this two-phase region, i.e., the information systems will resist merging and the integration effort will be very intensive and in some cases not resource efficient enough to pursue.

Emulsifiers are molecules with at least two active sites, one hydrophilic and the other hydrophobic. The hydrophobic end of the emulsifier attracts the non-polar molecules (such as oils) and the hydrophilic end attracts water. This enables hydrocarbons to dissolve in water. Depending on their structure and versatility, an ontology used to accomplish information integration could be compared to an emulsifier with multiple active sites.

15. Integration methodology

Data grouping [14] and axiom clustering [31] are important for both data and knowledge integration, respectively. Groups of similar data elements or axioms should be formed early in the integration. This is anal-

ogous to nucleation in liquids [12]. Grouping together similar entities in an information system can enhance integration efficiency.

Various dimensions of clustering depend on the clustering criteria, such as the formation of semantically heterogeneous groups [13,14], or grouping according to data categories [17]. Certain groups of data historically have been shown to exhibit more challenges to integration than others [17]. For example, administrative data such as date record loaded, date record changed, security classification and observation point etc. tend to be the data on which joins are based for application purposes [7]. Inconsistencies in these data will be noticed sooner than data that are used less frequently. Therefore, clusters involving this information should be formed. In an environment of limited resources, clusters in general should be selected to restrict the search domain to only those data elements and table names that are most likely to contain errors and inconsistencies. After information groups are formed, the integration should proceed at the ontological level.

Any good integration methodology will be able to handle special cases that arise due to anomalies in the information representation and content. These are not necessarily errors themselves, but rather they are conditions that could lead to errors. Ambiguous information representations can lead to erroneous integration that can interfere with the tractability (i.e., understanding the meaning of information). The information system analog of Fig. 4 illustrates how tractability can be higher in systems with components that are less integrated.

A liquid system of water and triethylamine has a lower consolute temperature because the constituents form a loosely bound compound that dissociates as the temperature is increased. The miscibility of water and triethylamine depends on the presence of this compound. Usually this is not a good model for information systems integration, Fig. 3 being the more likely case. However, some conditions in information systems are analogous to the phase diagram of water and triethylamine.

Using the same representation for what actually are disjoint domains can invite the wrong kind of query and lead to incorrect results that may look correct initially. For example, the abbreviations for distance units, nanometers and nautical miles, are both ''nm''. Using a database example, suppose exactly the same data representation for distance attributes were used in two tables, one of which described distances at sea and the other pertained to light wavelengths. Due to the apparent domain similarity, an erroneous join on distances could occur between a table that has ship speed data and a table describing the wavelengths of light from signals. The database management system would allow this meaningless join as legitimate unless additional software prevented it. The join results would be like the compound formed between water and triethylamine at lower temperatures. This apparent but false domain similarity occurs in many other cases in data standards where partially or totally disjoint domains are specified explicitly in the same format. In both cases, external factors serve as a context for the ''reaction'' or lack thereof.

16. Liquid crystals and long-ranger order

Liquid crystals are intermediate between liquids and crystalline solids [20,25,42,43]. Liquid crystals are materials consisting of anisotropic molecules. These materials exhibit some characteristics of liquids and some of solids [43]. Some researchers believe that liquid crystals represent a distinct state of matter that differs from crystalline solids and isotropic liquids [43]. Liquid crystals are substances that have long-range order in one or more physical dimensions, and only short-range order in the remaining dimensions. For example, *nematic* liquid crystals consist of long molecules, the major axes of which are oriented in about the same direction throughout a macroscopic domain, unlike an isotropic liquid in which the orientations of the molecules are not well correlated.

Similarly, *smectic* liquid crystals [25] consist of molecules that exhibit not only long-range order with respect to the orientations of the major molecular axes, but the molecular centers of mass are coplanar in a given domain [43]. However, the position of each of molecules in one plane with respect to the molecules in next plane is not correlated (ignoring average interplanar distance) and the layers can shear. Smectic liquid crystals have structures that bear quite a bit of similarity to three-dimensional solids. However, a smectic liquid crystal can be poured from one container to the next.

Other phases, such as a two-dimensional solid hexatic phase, as well as phase transitions such as two-dimensional melting have been observed. (See, for example, [20].)

The transition between liquid and solid corresponds to the transition from knowledge base to model base states of information. A liquid crystal is analogous to a knowledge base with many microtheories, each of which could be considered to be a model. As microtheories and domains of knowledge in knowledge bases become more refined with the right kind of detailed knowledge, a knowledge base of this type can become a de-facto model base.

17. Limitations of the methodology

First, analogies cannot be used to prove that any particular information system works better than any other one. The main purpose of analogy in this context is to suggest new ways to view, measure, and characterize information systems and to teach students about them. The use of analogy in general is not intended to be a rigorous form of scientific inquiry in the absence of other methods of investigation.

Second, the analogy between states of matter and states of information is expected to break down at some point. For example, T and P are well-known independent variables in a gas system of variable volume and fixed number of molecules. However, their infodynamic analogs, T_{DB} and E, are not nearly as well defined and are not independent of each other in the same sense that T and P are independent. The analogy also breaks down when one considers scalability issues.

For example, T and P are intrinsic variables, whereas T_{DB} and E are extrinsic because T_{DB} can decrease and E can increase with the size of the information system. We have no information system in which the number of data elements, axioms, or models comes anywhere near Avogadro's number. Information systems are already pushing the limits of tractability for $N \ll 10^{23}$. So far, no one has demonstrated the database analog of Avogadro's number, A_{DB}, has any particular significance, physical or otherwise.

A fundamental way in which matter and information differ is in their conservation and transfer (see Section 6). Like energy, matter is conserved whereas information is not. When matter is transferred from one location to another, there is a decrease in material in the former location and a corresponding increase in the final location. However, information can be transferred without any loss of information at the origin of the transfer.

Ultimately, the limitations of the analogy must be tested experimentally. Again, it suffices for purposes of discovery that the analogies are, at best, of a heuristic nature [33–35].

18. Future research and applications

More work is needed in this area to answer many questions. First, will k_{DB} be constant for all databases? Secondly, if not, will the range of k_{DB} be bounded in a predictable manner? How does one develop appropriate metrics for tractability (T_{DB}) and expressiveness (E)? Metrics techniques for knowledge bases have been the subject of a study in the now-concluded DARPA High Performance Knowledge Base Project (see, for example, [22]). This work continues today in the follow-on program, Rapid Knowledge Formation. It remains to be seen how much of these results can be applied to database systems.

Information grouping as compared to nucleation in matter needs to be explored further. Data grouping [14] in databases and axiom clustering [30,31] in knowledge bases are analogous to nucleation in gases and crystallization in liquids, respectively, because they initiate phase transitions to states with longer-range order and correlation among information entities. This is because these grouping techniques bring together data or knowledge in which the relationships between data elements or axioms link the elements together in the cluster or group in a manner analogous to the way in which intermolecular forces hold atoms or molecules together in condensed phases of matter. This area is fertile ground for further investigation.

Furthermore, model bases may be properly viewed as knowledge bases where the representational formalism has been extended in the form of a model. Just as the gas–liquid juncture becomes indeterminable above the critical temperature, the distinction between database, knowledge base, and model base may lack definition above some critical complexity of representation.

The above discussion described some ways to measure expressiveness (E) but metrics need to be developed for tractability. One possibility is to model tractability as the reciprocal of the time required to use or understand information in the system. More work is needed in this area.

The measurement of pressure seems quite trivial now, but this was not the case before the invention of the pressure gauge. Similarly, the measurement of expressiveness in information systems seems elusive now, but the future may prove otherwise.

More work is needed in the area of metrics for tractability and disjunction. The equations proposed in this paper should be tested and validated with use cases. A standard ontological representation needs to be established to enhance the value of disjunction metrics. Some liquid–liquid systems exhibit closed phase diagrams with both upper and lower consolute temperatures [3]. It may be of theoretical interest to determine if any information system exhibits analogous behavior and why. Solid–liquid equilibrium mixtures need to be explored as the information-system analog of model-base integration.

In addition to new database metrics, infodynamics principles can be used as the basis of a teaching method regarding the fundamentals of information systems for students already familiar with physical sciences.

By applying principles and properties of matter to information systems, scientists and engineers may be able to predict properties of future information systems in a manner that is analogous to the way in which we now predict the properties of future, undiscovered elements from knowledge of the periodic table. For example, model bases, when designed, developed, maintained, and managed efficiently, ought to provide an order of magnitude more modes of usage as an information system than either databases or knowledge bases.

19. Conclusion

Thermodynamics is but one domain from which we may draw analogical models for information systems. Pertaining to the mapping process itself, if we can use models to enable tractable computation, what about the tractability of the processes to find and verify those models? Clearly, one may proceed on an empirical basis – finding simple solutions, reusing them, and extending them as appropriate. That is to say that representation, including all processes of associative mapping, is evolutionary. This paper broadens one's perspective. For example, just as one may "borrow" from the chemical definition of simulated annealing in the formation of glasses and apply it to the optimization of neural networks, one also may borrow from the miscibility of two liquids based on their molecular polarity in the determination of segmentation in a knowledge base. The key is knowing when and where to apply the transformation(s). Such mappings may be seen as heuristic search, where the issue of representation is key. We believe that this paper has laid the foundation for associative mapping as an ontology in its own right. Then, ontologies can map other ontologies. The resulting network defines a randomization [19,33,35].

Acknowledgments

The authors thank the Defense Advanced Research Projects Agency, Office of Naval Research and the SSC-SD Science and Technology initiative for financial support. This work was produced by US government employees as part of their official duties and no copyright subsists therein. It is approved for public release with an unlimited distribution.

References

[1] P.N. Arora, On the Shannon measure of entropy, Information Sciences 23 (1) (1981) 1–9.
[2] G.W. Castellan, Physical Chemistry, Addison-Wesley Publishing Co., Reading, MA, 1964, pp. 11–13.
[3] G.W. Castellan, Physical Chemistry, Addison-Wesley Publishing Co., Reading, MA, 1964, pp. 291–296.
[4] G.W. Castellan, Physical Chemistry, Addison-Wesley Publishing Co., Reading MA, 1964, pp. 38–40.
[5] G.W. Castellan, Physical Chemistry, Addison-Wesley Publishing Co., Reading MA, 1964, pp. 29–34.
[6] G.W. Castellan, Physical Chemistry, Addison-Wesley Publishing Co., Reading MA, 1964, p. 75.
[7] M.G. Ceruti, (1) The Pressure Broadening of the Pure Rotational Raman Lines of Acetylene by Argon Gas; (2) The Pure Rotational Raman Spectrum of the Krypton Dimer; (3) A Fabry–Perot Interferometer Data Acquisition System. Ph.D. Dissertation, University of California at Los Angeles Department of Chemistry, 1979, pp. 11–12.
[8] M.G. Ceruti, (1) The Pressure Broadening of the Pure Rotational Raman Lines of Acetylene by Argon Gas; (2) The Pure Rotational Raman Spectrum of the Krypton Dimer; (3) A Fabry–Perot Interferometer Data Acquisition System. Ph.D. Dissertation, University of California at Los Angeles Department of Chemistry, 1979, p. 130.

[9] M.G. Ceruti, A review of data base system terminology, in: B.M. Thuraisingham (Ed.), Handbook of Data Management, Auerbach Publishers, Boca Raton, FL, 1998, p. 4, Chapter 1.
[10] M.G. Ceruti, An expanded review of information-system terminology. In: Proceedings of the AFCEA Federal Database Colloquium'99, 1999, pp. 173–191.
[11] M.G. Ceruti, States of matter and states of information. In: Proceedings of the ICSA 15th International Conference on Computer Applications in Industry and Engineering (CAINE 2002), San Diego, CA, November 2002, pp. 170–175.
[12] M.G. Ceruti, States of matter, information organization and dimensions of expressiveness. In: Proceedings of the 2004 ACM International Conference on Computing Frontiers (CF04), Ischia, Italy, April 2004, pp. 120–124.
[13] M.G. Ceruti, M.N. Kamel, Semantic heterogeneity in database and data dictionary integration for command and control systems. In: Proceedings of the DOD Database Colloquium'94, San Diego, CA, August 1994, pp. 65–89.
[14] M.G. Ceruti, M.N. Kamel, Preprocessing and integration of data from multiple sources for knowledge discovery, International Journal on Artificial Intelligence Tools 8 (2) (1999) 152–177.
[15] M.G. Ceruti, S.J. McCarthy, Establishing a data-mining environment for wartime event prediction with an object-oriented command and control database. In: Proceedings of the Third IEEE International Symposium on Object-oriented Real-time Distributed Computing, ISORC2K, Newport Beach CA, March 2000, IEEE Computer Society Press, 2000, pp. 174–179.
[16] M.G. Ceruti, S.H. Rubin, Infodynamics III: information integration and tractability. In: Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI-2005), Las Vegas, NV. August 15–17, 2005, pp. 25–31.
[17] M.G. Ceruti, B.M. Thuraisingham, M. N. Kamel, Restricting search domains to refine data analysis in semantic-conflict identification. In: Proceedings of the 17th AFCEA Federal Database Colloquium and Exposition'99, San Diego, CA, September 2000, pp. 211–218.
[18] M.G. Ceruti, T.L. Wright, Knowledge management for distributed tracking and the next-generation command and control. In: Proceedings of IEEE International Software Metrics Symposium (METRICS-2005), Industry track, Como Italy, September 19–22, 2005, pp. 1–4.
[19] G.J. Chaitin, Randomness and mathematical proof, Scientific American 232 (5) (1975) 47–52.
[20] C.F. Chou, A.J. Jin, S.W. Hui, C.C. Huang, J.T. Ho, Multiple-step melting in two-dimensional hexatic liquid-crystal films, Science 280 (5368) (1998) 1424–1426.
[21] E.F. Codd, A relational model of data for large shared data banks, Communications of the ACM 13 (6) (1970) 377–387. Available from: <http://www.acm.org/classics/nov95/toc.html>.
[22] P. Cohen, R. Schrag, E. Jones, A. Pease, A. Lin, B. Starr, D. Easter, D. Gunning, M. Burke, The DARPA high performance knowledge bases project, AI Magazine 19 (4) (1998) 25–49.
[23] A.H. Darooneh, Utility function from maximum entropy principle, Entropy, 1099-4300 8 (1) (2006) 18–24. Available from: <http://www.mdpi.org/entropy/>.
[24] D. Filev, R.R. Yager, Analytic properties of maximum entropy OWA operators, Information Sciences 85 (1–3) (1995) 11–27.
[25] J.W. Goodby, M.A. Waugh, S.M. Stein, E. Chin, R. Pindak, J.S. Patel, Characterization of a new helical smectic liquid crystal, Nature 337 (2) (1989) 452–499.
[26] A.J. Grove, J.Y. Halpern, D. Koller, Random worlds and maximum entropy, Journal of Artificial Intelligence Research 2 (1994) 33–88.
[27] S. Guha, N. Mishra, R. Motwani, L. O'Callaghan, Clustering in data streams, Proceedings of the IEEE Symposium on Foundations of Computer Science (November) (2000) 359–366.
[28] D.B. Lenat, R.V. Guha, Building Large Knowledge-based systems: Representation and Inference in the Cyc Project, Addison-Wesley, Reading, MA, 1989, 372 pp.
[29] H. J. Levesque, R. J. Brachman, A fundamental tradeoff in knowledge representation and reasoning (revised version); The knowledge representation enterprise, pp. 41–70 (Chapter 4); Original version appeared as H. J. Levesque, A Fundamental Tradeoff in Knowledge Representation and Reasoning. In: Proc of CSCSI/SCEIO Conference (CSCSI-84), London, Ontario, 1984, pp. 141–152.
[30] M. Mehrotra, Ontology analysis for the semantic web. In: Proceedings of the AAAI-02/IJCAI-02 Workshop on Ontologies and the Semantic Web, Technical Report WS-02-11, Edmonton, Alta., Canada, 2002, pp. 41–51.
[31] M. Mehrotra, C. Wild, Analyzing knowledge-based systems using multi-viewpoint clustering analysis, Journal of Systems and Software 29 (3) (1995) 235–249.
[32] W. Rödder, C.-H. Meyer, Coherent knowledge processing at maximum entropy by SPIRIT. In: Proceedings of UAI-96, Morgan Kaufmann, 1996, pp. 470–476.
[33] S.H. Rubin, On the auto-randomization of knowledge. In: Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, Las Vegas NV, 2004, pp. 308–313.
[34] S.H. Rubin, M.G. Ceruti, R.J. Rush Jr. Knowledge mining for decision support in command and control systems. In: Proceedings of the 17thAFCEA Federal Database Colloquium and Exposition'99, 2000, pp. 127–133.
[35] S.H. Rubin, S.N.J. Murthy, M.H. Smith, L. Trajkovic, KASER: knowledge amplification by structured expert randomization, IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 34 (6) (2004) 2317–2329.
[36] H.B. Woolf, Editor in Chief, Webster's New Collegiate Dictionary, G.&C. Merriam Co. Springfield, MA, 1979, p. 470.
[37] H.B. Woolf, Editor in Chief, Webster's New Collegiate Dictionary, G.&C. Merriam Co. Springfield, MA, 1979. p. 665.
[38] H.B. Woolf, Editor in Chief, Webster's New Collegiate Dictionary, G.&C. Merriam Co. Springfield, MA, 1979. p. 1098.
[39] H.B. Woolf, Editor in Chief, Webster's New Collegiate Dictionary, G.&C. Merriam Co. Springfield, MA, 1979. p. 401.
[40] R.R. Yager, Measures of entropy and fuzziness related to aggregation operators, Information Sciences 82 (3–4) (1995) 147–166.
[41] W.H. Zurek (Ed.), Complexity, Entropy and the Physics of Information, Addison-Wesley, New York, 1991.

[42] Available from: <http://bly.colorado.edu/lc/lcintro.html>.
[43] Available from: <http://abalone.cwru.edu/tutorial/enhanced/files/lc/phase/phase.htm>.

Dr. Marion G. Ceruti is a senior scientist in the Command and Control Technology and Experimentation Division of the Command and Control Department at the Space and Naval Warfare Systems Center, San Diego. She received the Ph.D. in chemistry in 1979 from the University of California at Los Angeles. Dr. Ceruti's professional activities include information systems research and analysis for command and control decision-support systems, sensor fusion, information support for chemical research and research management. She is the author of 95 journal articles, conference proceedings, monographs and book chapters on various topics in science and engineering. She received several publication awards, including the SSC San Diego Publication of the Year for her first paper on Infodynamics [11].Dr. Ceruti is a senior member of the IEEE and a member of the IEEE Computer Society, the American Chemical Society, the Association for Computing Machinery, the Armed Forces Communications and Electronics Association (AFCEA), the International Society for Computers and Their Applications, and the New York Academy of Sciences. She was the program chairperson for the AFCEA Federal Database Colloquium for 14 years and has served on program and local-arrangements committees for IEEE conferences.

Dr. Stuart H. Rubin is a senior scientist at the Space and Naval Warfare Systems Center (SSC) in San Diego, code 2734. He received a BS in business from the University of Rhode Island, Kingston, RI in 1975; an MS in industrial and systems engineering from Ohio University in Athens, OH in 1977, an MS in computer science from Rutgers University in Piscataway, NJ in 1980, and a Ph.D. in computer and information science from Lehigh University in Bethlehem, PA in 1988. In 2003, he was awarded the IEEE Systems, Man, and Cybernetics Society's Outstanding Service Award.

Dr. Rubin has authored over 150-refereed conference and journal papers as well as several patent applications on behalf of SSC, San Diego. His professional interests center about heuristic methodologies for multi-sensor fusion, associative memory, decision support, and knowledge discovery.

Dr. Rubin is a senior member of IEEE and a member of the AFCEA, the American Association for the Advancement of Science, the New York Academy of Sciences, the North American Fuzzy Information Processing Society, and several other scientific societies. Dr. Rubin chairs the IEEE Systems, Man, and Cybernetics Technical Committee on knowledge acquisition in intelligent systems. Dr. Rubin is also an associate editor of the IEEE Transactions on Systems, Man, and Cybernetics, Part C as well as the International Journal of Modeling and Simulation and the Journal of Systemics, Cybernetics, and Informatics. Dr. Rubin is the founder and general co-chair of the IEEE International Conference on Information Reuse and Integration (IRI); and, has delivered several keynote lectures at international conferences. Dr. Rubin was previously a member of the SMC Adcom and currently serves on the SMC Board of Governors (BoG).